CSE 251A Winter 2021 - Take Home Exam 3

1 hour **Name:** **PID:**

---

(1) This is an open book, open notes exam. You are free to consult any text book or notes. **You are not allowed to consult with any other person.**
(2) If you need any clarification, please post a private message to the instructors on Piazza.
(3) Remember that your work is graded on the *clarity* of your writing and explanation as well as the validity of what you write.
(4) This is a one-hour exam.

(1) In class we looked at convex and strongly convex functions. A function $f : \mathbb{R}^d \to \mathbb{R}$ is said to be positively convex if one of the following three conditions hold:
   (a) For all $x, y \in \mathbb{R}^d$ and $0 < t < 1$, $f(tx + (1-t)y) < tf(x) + (1-t)f(y)$.
   (b) If f is differentiable then for all $x, y \in \mathbb{R}^d$, $f(y) > f(x) + \nabla f(x)^\top (y - x)$.
   (c) If f is doubly differentiable, then for all $x \in \mathbb{R}^d$, $\nabla^2 f(x)$ is positive definite. In other words for all $z \in \mathbb{R}^d$ such that $z \neq 0$, we have $z^\top \nabla^2 f(x) z > 0$.

   Notice that the difference between convexity and positive convexity is that the inequalities are strict. In what follows, w and x are vectors in $\mathbb{R}^d$ and y is a scalar. Answer the following questions.

   (a) (5 points) Is the function $f(w) = \log(1+e^{-yw^\top x})$ a positively convex function of w? Justify your answer.
   Solution: We need to find $\nabla^2 f(w)$, and see if it is positive semi definite or positive definite. We first find the gradient of $f(w)$ with respect to w:

$$\frac{d \log(1 + e^{-yw^\top x})}{dw} = -yx\left(\frac{e^{-yw^\top x}}{(1 + e^{-yw^\top x})}\right)$$

   To help simplify the notation, we can define a function $g(z) = \frac{1}{(1+e^{-z})}$. Therefore, $1 - g(z) = \frac{e^{-z}}{(1+e^{-z})}$. With this new notation, we can write:

$$\nabla f(w) = -yx(1 - g(yw^\top x)) = -yx + yxg(yw^\top x)$$

   Now, we need to take the second derivative. Since the first term doesn't have w in it, we can ignore it and for the second term, we can write:

$$\nabla^2 f(w) = \frac{d \quad yxg(yw^\top x)}{d \quad w} = y^2 xx^\top \cdot g'(yw^\top x) = y^2 xx^\top \cdot \frac{e^{-yw^\top x}}{(1 + e^{-yw^\top x})^2}$$

   y^2 and $\frac{e^{-yw^\top x}}{(1+e^{-yw^\top x})^2}$ are both non-negative scalars, so we only need to look at the matrix $xx^\top$ for checking whether the function is positively convex or only convex. We know that xx^t is PSD (as showed in the lectures and also the discussion section), but to make sure it is not PD we need to find a vector v such that $v^\top xx^\top v = 0$. Let's pick a vector v s.t. $||v|| = 1$ and $< v, x >= 0$, in other words v is a vector orthogonal to x. Then, $v^\top xx^\top v = (v^\top x)^2 = 0$ as well, which shows that $xx^\top$ is only PSD and not PD. Thus, the function is only convex and not positively convex.

   (b) (5 points) Let $g(w) = (w^\top x - y)^2 + \frac{1}{2}\|w\|^2$. Is g a positively convex function of w? Justify your answer.
   Solution: We need to find $\nabla^2 g(w)$, and see if it is positive semi definite or positive definite. We first find the gradient of $g(w)$ with respect to w:

$$\nabla g(w) = 2x(w^\top x - y) + \frac{1}{2} \cdot 2w = 2xw^\top x - 2xy + w$$

Now, we differentiate once more, to find $\nabla^2 g(w)$. Since $-2xy$ doesn't have w in it, the derivative is all zeros for all dimensions. The derivative of w with respect to w is the identity matrix, I_d. So, now we only need to find the derivative of $2xw^\top x$:

$$\frac{d(2xw^\top x)}{dw} = \frac{d(2x(w^\top x))}{dw} = \frac{d(2x(x^\top w))}{dw} = \frac{d(2(xx^\top)w)}{dw} = 2xx^\top$$

Therefore, the Hessian of $g(w)$ is:

$$\nabla^2 g(w) = 2xx^\top + I_d$$

We've shown in the lectures and also the discussion sections that the matrix $xx^\top$ is PSD. We also showed that the identity matrix I_d is PD. The addition of a PSD and PD matrix is PD. We can show this by writing $v^\top(2xx^\top + I_d)v = v^\top(2xx^\top)v + v^\top(I_d)v = 2 < x, v >^2 + < v, v >= 2 < x, v >^2 + ||v||^2$, where $< x, v >^2$ is non-negative and can take the value zero, but $||v||^2 > 0$, since v is non-zero. Therefore, the summation is strictly larger than zero and $\nabla^2 g(w)$ is PD, which means g is positively convex.

(2) We have a training set $S = \{(x^{(i)}, y^{(i)}), i = 1, \ldots, n\}$. Suppose we transform the feature vectors by scaling each vector by a constant $c > 0$ to get a new training set T. In other words, $T = \{(cx^{(i)}, y^{(i)}, i = 1, \ldots, n\}$. Assume that we are looking for linear classifiers whose decision boundary pass through the origin – namely, the term $b = 0$.

(a) (5 points) Write down the Perceptron algorithm for training set T, with the weight vector initialized as $w = 0$.

Solution: For this simple classifier, for sample i, if $y^{(i)}(w_T c x^{(i)}) > 0$, then the loss is zero. However, if $y^{(i)}(w_T c x^{(i)}) \leq 0$, the loss is $-y^{(i)}(w_T c x^{(i)})$. Given this, we can write the algorithm as follows:

(i) input: dataset T

(ii) Initialize the weights vector $w_{T_0} = 0$

(iii) For i in T:
if $y^{(i)}(w_{T_t} c x^{(i)}) \leq 0$:

$$w_{T_{t+1}} = w_{T_t} + y^{(i)} c x^{(i)}$$

(iv) output: vector w_T

(b) (5 points) Suppose w_S is the output of Perceptron for S and w_T is the output of Perception for T. Is $w_S = w_T$? Justify your answer through a brief proof or counterexample.

Solution: Let's assume we apply the Perceptron algorithm for both datasets S and T, and iterate through the points from $i = 1$ to $i = n$, in the same order. Since we initialize both w_S and w_T with zero, then for both datasets, we either need to do an update (we had a mis-classification) or we don't. If we do apply an update, for S, it will be $w_{S_1} = w_{S_0} + y^{(1)}x^{(1)} = y^{(1)}x^{(1)}$ and for T it will be $w_{T_1} = w_{T_0} + y^{(1)}cx^{(1)} = cy^{(1)}x^{(1)}$. We can see that $w_{T_1} = c \cdot w_{S_1}$. For the second sample, since the weights only differ by the positive coefficient c, the decision produced by the classifiers would be the same, and we either have to apply an update for both cases, or we don't. Given this, once the algorithm converges for the datasets, we will have $w_S = \sum_{i \in U} y^{(i)}x^{(i)}$ and $w_T = c\sum_{i \in U} y^{(i)}x^{(i)}$, where U is the set of samples for which we need to do an update. Given this, we see that $w_T = cw_S$.